$t_{14,0.02} = 2.264$

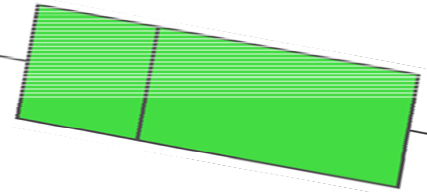
$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{(x-\mu)^2 \mid 2\sigma^2\}$

$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$

$\bar{x} = \frac{1}{n}\sum x_i$

$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

1	02667
2	22389
3	45
4	58
5	9
6	11



Chapter 25: Two Categorical Variables: The Chi-Square test

Recall – Two way tables

- An experiment has a **two-way** design if two **categorical** factors are studied with several levels of each factor.

- Two-way tables organize data about two categorical variables with any number of levels/treatments obtained from a two-way, or block, design.

MARGINAL AND CONDITIONAL DISTRIBUTIONS

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

A **conditional distribution** of a variable is the distribution of values of that variable among only individuals who have a given value of the other variable. There is a separate conditional distribution for each value of the other variable.

A financial aid officer is studying the relationship between who borrows money for college in a family and the income of the family. Is there evidence that family income is associated with who borrows money for college, the parent(s) or the student?

Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202	45	55	200
$65k – $250k	240	119	170	250
> $250k	115	55	47	250
Total	557	219	272	700

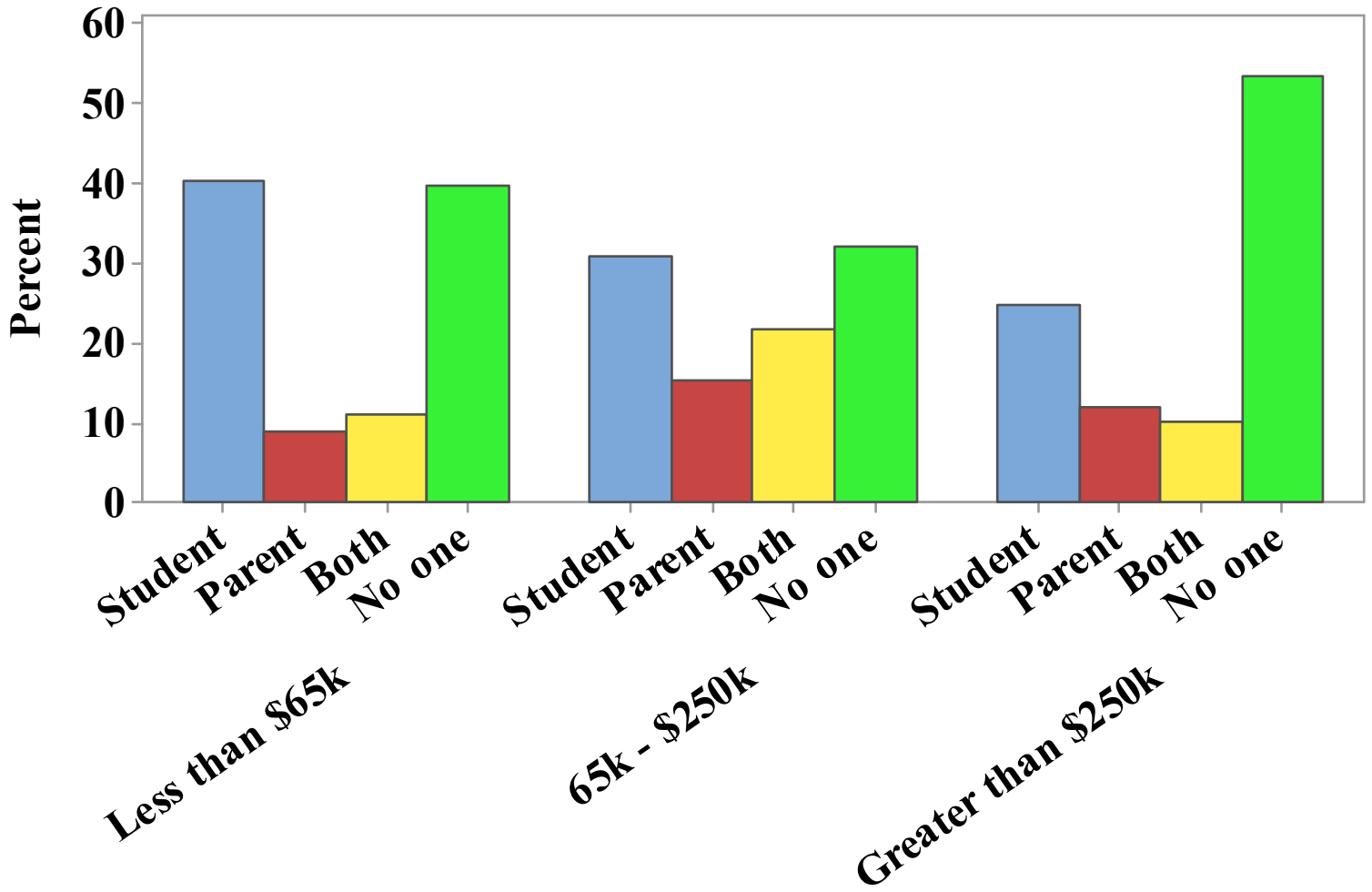
A financial aid officer is studying the relationship between who borrows money for college in a family and the income of the family. Is there evidence that family income is associated with who borrows money for college, the parent(s) or the student?

RESPONSE VARIABLE:

EXPLANATORY VARIABLE

Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202 40.24%	45 8.96%	55 10.96%	200 39.84%
$65k – $250k	240 30.81%	119 15.28%	170 21.82%	250 32.09
> $250k	115 24.63%	55 11.78%	47 10.06%	250 53.53%
Total	557	219	272	700

- Are the true conditional distributions of who borrows the same for these three income groups?



- What can we say if the true conditional distributions are the same for each income group?

Inference about associations

Null Statement: There is no relationship between categorical variable A and categorical variable B.

Alternative Statement: There is some relationship between categorical variable A and categorical variable B.

The steps for multiple comparisons tests

Statistical methods with many possible alternatives have two steps:

1. An overall test to see if there is good evidence of any differences among parameters that we want to compare.
2. When the overall test is significant, a detailed follow up analysis is used to decide which of the parameters differ and to estimate how large the differences are.

Expected counts in two-way tables

We want to test the hypothesis that there is no difference in the conditional distribution of the response across levels of the explanatory (H_0).

To test this hypothesis, we compare actual counts from the sample data to the expected counts we would observe if the null hypothesis was true.

Under H_0 , the expected count in a cell of a two-way table is calculated as:

Calculate expected counts assuming the null is true

Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202	45	55	200
$65k – $250k	240	119	170	250
> $250k	115	55	47	250
Total	557	219	272	700

The Basic Test Approach

- Assume the conditional distribution of the response is the same for each level of the explanatory variable.
  - Assume there is no relationship and calculate how many observations we would expected to be in each cell of the table
- Compare the observed counts to these expected counts.
- Suppose the difference between the observed and the expected counts is small:
- Suppose the difference between the observed and the expected counts is large:

The chi-square test

- Are the differences we see between the observed and expected counts likely to have occurred just by chance because of the random sampling?
- The chi-square test statistic (χ^2) is a measure of how much the observed cell counts differ from the expected cell counts.
- A large difference between the observed and expected counts is evidence against the null hypothesis of no difference in the conditional distributions.

$$\chi^2 \text{component} = \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

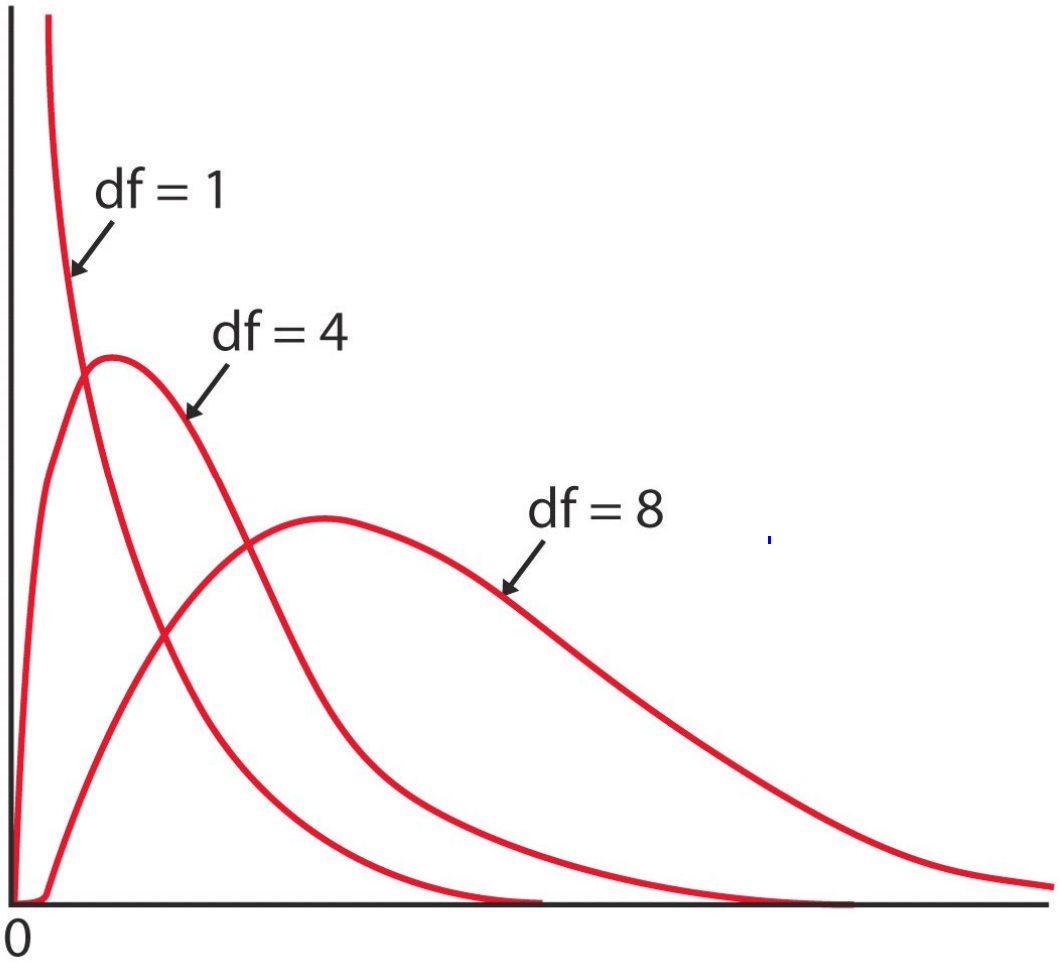
Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202 160.0	45 62.9	55 78.1	200 201.0
$65k – $250k	240 248.2	119 97.6	170 121.2	250 312.0
> $250k	115 148.8	55 58.5	47 72.7	250 187.0

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

The chi-square distributions

The chi-square distributions are a family of distributions that can take only positive values, are skewed to the right, and are described by specific degrees of freedom.

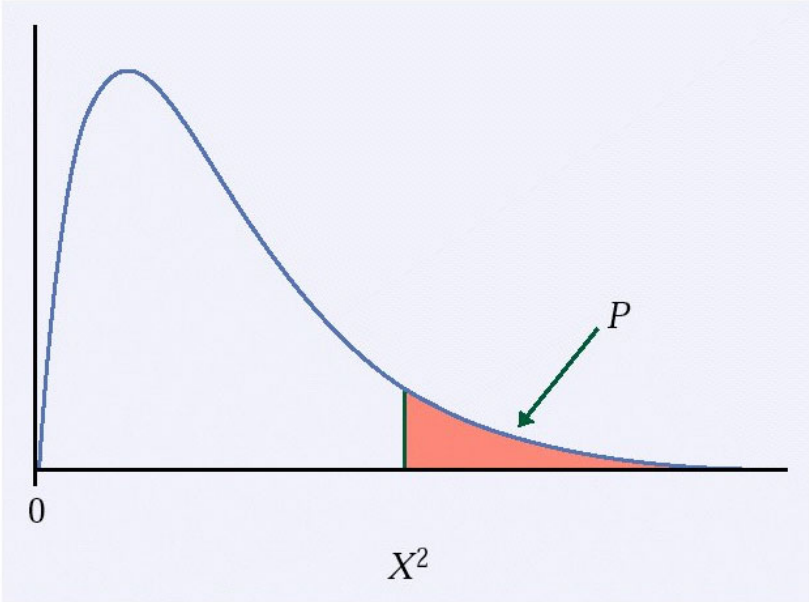
Table D gives upper critical values for many chi-square distributions.



For the chi-square test, H_0 states that there is no difference in the conditional distributions. The alternative is that there is a difference in the conditional distributions.

Under H_0, the chi-square test has approximately a χ^2 distribution with

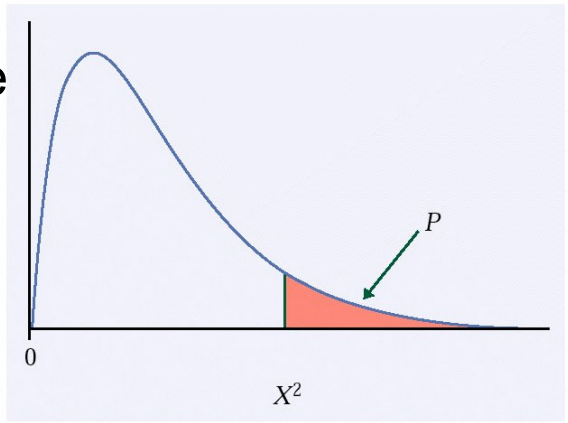
---



The *p*-value for the chi-square test is the area to the right of χ^2 :

$$P\left[\chi^2 \geq x^2\right]$$

Find the p-value



Chi-square distribution critical values

	Upper tail probability p			
df	.005	.0025	.001	.0005
3	12.84	14.32	16.27	17.73
4	14.86	16.42	18.47	20.00
5	16.75	18.39	20.51	22.11
6	18.55	20.25	22.46	24.10
7	20.28	22.04	24.32	26.02

Conclusion in the context of the problem (part 1 – the overall test result)

-
-
-

$t_{14,0.02} = 2.264$

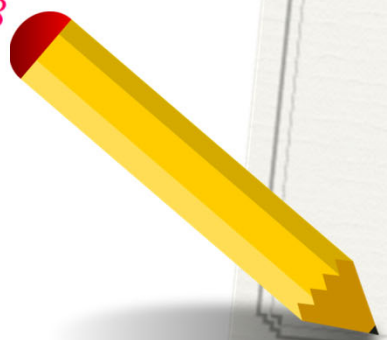
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{(x-\mu)^2 \mid 2\sigma^2\}$$

$$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

1	02667
2	22389
3	45
4	58
5	9
6	11



Chapter 25: Two categorical variables: The Chi-Square test

A financial aid officer is studying the relationship between who borrows money for college in a family and the income of the family. Is there evidence that family income is associated with who borrows money for college, the parent(s) or the student?

RESPONSE VARIABLE:

EXPLANATORY VARIABLE

Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202 40.24%	45 8.96%	55 10.96%	200 39.84%
$65k – $250k	240 30.81%	119 15.28%	170 21.82%	250 32.09
> $250k	115 24.63%	55 11.78%	47 10.06%	250 53.53%
Total	557	219	272	700

Chi-square test = homogeneity of proportions

Q: Are the true conditional distributions of who borrows money the same for each income group, or do they differ??

The steps for multiple comparisons tests

Statistical methods with many possible alternatives have two steps:

1. An overall test to see if there is good evidence of any differences among parameters that we want to compare.
2. When the overall test is significant, a detailed follow up analysis is used to decide which of the parameters differ and to estimate how large the differences are.

We have a significant chi-square result

- Which factor combination(s) contributed the most to the chi-square test statistic?

Family Income	Who borrows money			
	Student only	Parent only	Both	No one
< $65k	202 11.05 160.0	45 5.09 62.9	55 6.84 78.1	200 0.01 201.0
$65k – $250k	240 0.27 248.2	119 4.69 97.6	170 19.63 121.2	250 12.31 312.0
> $250k	115 7.68 148.8	55 0.21 58.5	47 9.07 72.7	250 21.21 187.0

Conclusion in the context of the problem (part 2 – the nature of the relationship)

-
-
-

CELL COUNTS REQUIRED FOR THE CHI-SQUARE TEST

You can safely use the chi-square test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, all four expected counts in a 2×2 table should be 5 or greater.